Autonomous Zero-Shot 6D Pose Estimation Pipeline

Project Overview

This project demonstrates a zero-shot, open-world pipeline for estimating the 6D pose of objects in RGB images without requiring pre-collected CAD models. It combines semantic reasoning (Moondream2 & Gemini) with automated geometry retrieval (Objaverse) and mesh processing for robotics perception.

Pipeline Components

1. RGB Image Input
   - Any standard RGB image containing objects.
   - Supports single images for demonstration or batch processing.

2. Scene & Object Extraction (Moondream2)
   - Uses a vision-language model to describe the scene in 3-5 words.
   - Lists individual objects present in the image.

3. LVIS Category Matching
   - Maps open-vocabulary object names to LVIS categories.
   - Uses Gemini for context-aware relabeling if local match fails.

4. 3D Mesh Retrieval (Objaverse)
   - Automatically downloads meshes corresponding to the LVIS category.
   - Supports multiple candidates per object for redundancy.

5. Mesh Decimation
   - Uses fast_simplification to reduce mesh vertices.
   - Ensures deployment-ready meshes (≤ 50k vertices).

6. Export & Integration
   - Saves decimated OBJ meshes to the configured output directory.
   - Compatible with FoundationPose for 6D pose estimation.

Configuration

1. Install Python dependencies:
   pip install torch transformers trimesh objaverse fast_simplification
   google-generativeai pillow reportlab

2. Set environment / API keys:
   - Gemini API key (for semantic relabeling)
   - Ensure network access to Objaverse

3. Configure constants in `zero_shot_mesh_pipeline.py`:
   - IMAGE_PATH : Path to input image
   - OUTPUT_DIR : Directory to save meshes
   - MAX_MESHES_PER_LABEL : Max meshes per object
   - MAX_VERTICES : Target vertex limit